W130–W134 *Nucleic Acids Research, 2004, Vol. 32, Web Server issue*
DOI: 10.1093/nar/gkh366

siRNA Selection Server: an automated siRNA oligonucleotide prediction server

Bingbing Yuan, Robert Latek, Markus Hossbach[1], Thomas Tuschl[2] and Fran Lewitter*

Whitehead Institute for Biomedical Research, Bioinformatics and Research Computing, Nine Cambridge Center, Cambridge, MA 02142, USA, [1]Max Planck Institute for Biophysical Chemistry, Göttingen, Germany and [2]Laboratory of RNA Molecular Biology, Rockefeller University, NY 10021, USA

Received February 15, 2004; Revised and Accepted March 5, 2004

ABSTRACT

The Whitehead siRNA (short interfering RNA) Selection Web Server (http://jura.wi.mit.edu/bioc/siRNA) automates the design of short oligonucleotides that can specifically 'knock down' expression of target genes. These short sequences are about 21 nt in length, and when synthesized as double stranded RNA and introduced into cell culture, can reduce or eliminate the function of the target gene. Depending on the length of a gene, there are potentially numerous combinations of possible 21mers. Some experimental evidence has already shown that not all 21mers in a gene have the same effectiveness at silencing gene function. Our tool incorporates published design rules and presents the scientist with information about uniqueness of the 21mers within the genome, thermodynamic stability of the double stranded RNA duplex, GC content, presence of SNPs and other features that may contribute to the effectiveness of a siRNA.

INTRODUCTION

One way to study the function of a gene is to reduce or eliminate its expression in a cell. An emerging technology to study the role of an individual gene is RNAi and its effector molecule siRNA—short interfering RNA (1,2). A properly selected short double stranded RNA (~21 nt) targeted to a specific sequence can silence the expression of the gene. Some experimental evidence has already shown that not all 21mers in a gene have the same effectiveness at silencing gene function [see, for example (1,3–6)]. We have used a number of bioinformatics approaches to help select optimal siRNAs. The computational tools we build are useful to scientists wishing to silence specific genes, specific gene families, or even specific biological pathways within a genome.

In collaboration with laboratory scientists at Whitehead Institute and the Max Planck Institute we have built a web-based tool for siRNA selection (http://jura.wi.mit.edu/bioc/siRNA) that implements several algorithms to identify siRNAs with a high probability of silencing the target gene. This server has been available since November 2002. As new experimental results are reported, we incorporate these rules into our website so that researchers can easily get access to the new design features. Our website provides the biologist with the flexibility to use predefined siRNA patterns or input their own patterns. Several filters can refine users' oligonucleotide sequence characteristics, such as GC percentage, base variations and the number of repetitive bases in a row. Since the objective of using siRNA is to silence a specific gene in a mammalian cell, the base-pairing region for a siRNA is carefully selected to avoid similarity to any unrelated mRNA. To do so, our program incorporates similarity searching of each candidate siRNA against the human or mouse UniGene databases. Subsequently, each candidate siRNA is mapped to the human or mouse genome, indicating if the siRNA maps to an exon–exon boundary. To aid in the selection of a siRNA from a region of minimal genetic variation, published single nucleotide polymorphisms (SNPs) in the region of each candidate siRNA are also shown.

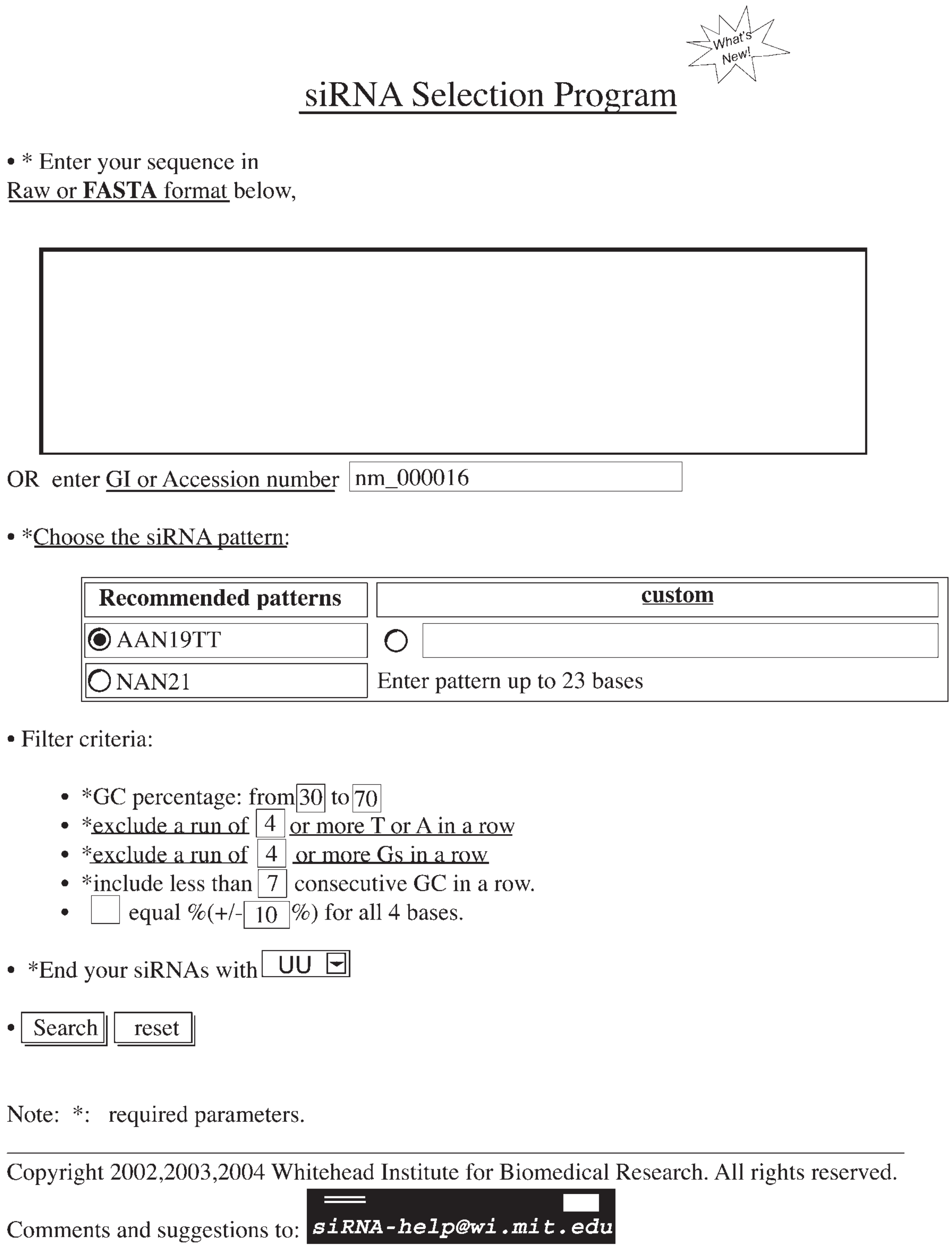
ACCESS TO SERVER

In order to access our server, we require user registration. This permits us to limit the number of searches per day for an individual investigator. Our current limit is 15 searches per day. The use of this site is provided free of charge to the research community.

INPUT TO THE SERVER

Figure 1 shows the input form that guides the biologist through the design of siRNAs. You can enter either a cDNA sequence in raw or FASTA format, or a GI or GenBank accession

*To whom correspondence should be addressed. Tel: +1 617 258 5000; Fax: +1 617 258 5578; Email: sirna-help@wi.mit.edu



siRNA Selection Program

- * Enter your sequence in Raw or **FASTA** format below,

OR enter GI or Accession number nm_000016

- *Choose the siRNA pattern:

Recommended patterns	**custom**
● AAN19TT	○
○ NAN21	Enter pattern up to 23 bases

- Filter criteria:
  - *GC percentage: from 30 to 70
  - *exclude a run of 4 or more T or A in a row
  - *exclude a run of 4 or more Gs in a row
  - *include less than 7 consecutive GC in a row.
  - □ equal %(+/- 10 %) for all 4 bases.
- *End your siRNAs with UU
- Search reset

Note: *: required parameters.

Copyright 2002,2003,2004 Whitehead Institute for Biomedical Research. All rights reserved.

Comments and suggestions to: siRNA-help@wi.mit.edu

Figure 1. Data input web page to begin design of siRNAs for your target sequence.

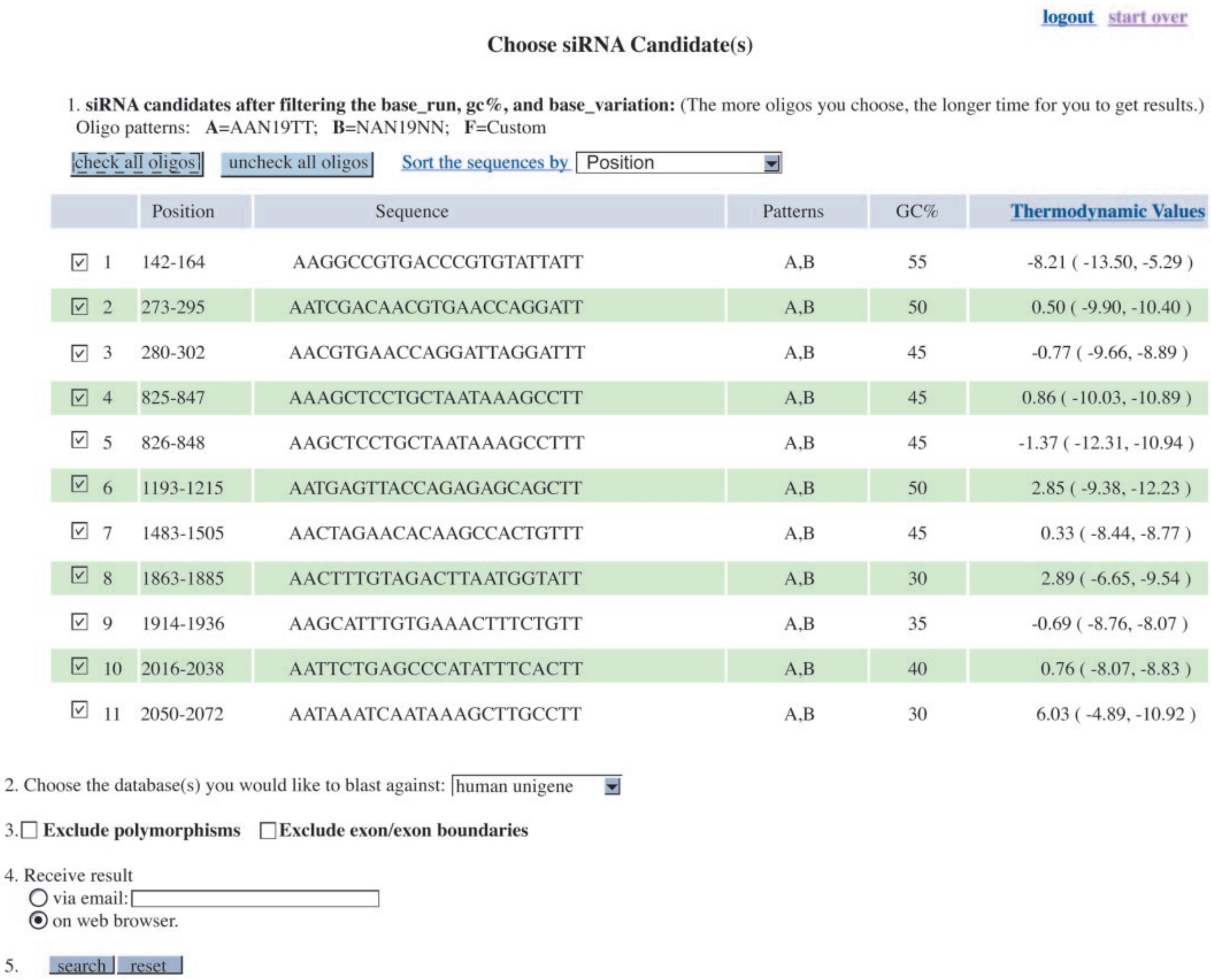
number. Next you need to choose a pattern for the composition of the siRNA. There are several commonly used siRNA sequence pattern (for example, AAN19TT; for more detail see http://www.rockefeller.edu/labheads/tuschl/sirna.html). Our website provides biologists the flexibility to use these predefined siRNA patterns or input their own pattern. The full NC-IUB nomenclature can be used to represent the sequence. In addition, a modified regular expression pattern can be constructed.

Several filters related to base composition are also available in the design process. Sequence characteristics such as GC percentage, base variations and the number of repetitive bases in a row are available for selection. These filters are based on experimental results observed by Tuschl and others. A run of four or more Ts or As should be excluded under some circumstances because four or five Ts in a row is the transcription terminator signal for pol III. If it is desired to design hairpin RNA expression vectors that are expressed from pol III promoters (U6, H1, or tRNA promoter), pol III terminator signals must be excluded from the sense or anti-sense strand. Similarly four or more Gs in a row should be excluded because oligoG-containing RNAs may form tetraplexes and are difficult to chemically synthesize with some, but not all, types of RNA chemistry. GC rich sequences form more stable duplexes than those that are AT rich, thus more than seven G/C pairs in a row would be suboptimal.

The final step in design is to select the terminal bases for the siRNA. To construct the siRNAs, we first find all 23mers with your pattern. For the sense strand, we consider the 21mer from positions 3 to 23 of the candidate siRNAs for further analysis. For the anti-sense strand, the 21mer is the reverse complement of positions 1 to 21 of the sense strand. The two bases at the 3′ end of the 21mers are replaced by your choice of terminal bases. If you choose NN, the original two nucleotides are kept intact.

SELECTING OLIGOS FOR FURTHER ANALYSIS

The next screen (Figure 2) shows the potential siRNAs based on your input sequence, pattern and filtering choices. Although the hits are initially sorted by position within the target sequence, the results can be sorted by various criteria. The interesting information includes the position of the siRNA within the input sequence, the pattern the siRNA matches, the percentage GC content and the thermodynamic values based on stability of the 5′ ends of the sense and anti-sense siRNAs. Recent experimental evidence indicates that the two strands of a siRNA duplex do not enter into the RNAi pathway equally. Rather, the less stable 5′ end (either sense or anti-sense) in the siRNA duplex directs that strand to enter into the RISC complex and has more effect silencing the target gene (7–10). To calculate stability of the duplex, we examine the free energy at its two ends using the nearest neighbor method (11). For each siRNA, a model helix is made from each end of the duplex with five bases of the 5′ end of the first strand, four Ns, then seven bases of the 3′ end of the second strand. For each model, thermodynamic parameters for four nearest



Choose siRNA Candidate(s)

1. **siRNA candidates after filtering the base_run, gc%, and base_variation:** (The more oligos you choose, the longer time for you to get results.)
Oligo patterns: **A**=AAN19TT; **B**=NAN19NN; **F**=Custom

Sort the sequences by: Position

		Position	Sequence	Patterns	GC%	Thermodynamic Values
☑	1	142-164	AAGGCCGTGACCCGTGTATTATT	A,B	55	-8.21 (-13.50, -5.29)
☑	2	273-295	AATCGACAACGTGAACCAGGATT	A,B	50	0.50 (-9.90, -10.40)
☑	3	280-302	AACGTGAACCAGGATTAGGATTT	A,B	45	-0.77 (-9.66, -8.89)
☑	4	825-847	AAAGCTCCTGCTAATAAAGCCTT	A,B	45	0.86 (-10.03, -10.89)
☑	5	826-848	AAGCTCCTGCTAATAAAGCCTTT	A,B	45	-1.37 (-12.31, -10.94)
☑	6	1193-1215	AATGAGTTACCAGAGAGCAGCTT	A,B	50	2.85 (-9.38, -12.23)
☑	7	1483-1505	AACTAGAACACAAGCCACTGTTT	A,B	45	0.33 (-8.44, -8.77)
☑	8	1863-1885	AACTTTGTAGACTTAATGGTATT	A,B	30	2.89 (-6.65, -9.54)
☑	9	1914-1936	AAGCATTTGTGAAACTTTCTGTT	A,B	35	-0.69 (-8.76, -8.07)
☑	10	2016-2038	AATTCTGAGCCCATATTTCACTT	A,B	40	0.76 (-8.07, -8.83)
☑	11	2050-2072	AATAAATCAATAAAGCTTGCCTT	A,B	30	6.03 (-4.89, -10.92)

2. Choose the database(s) you would like to blast against: human unigene

3. ☐ **Exclude polymorphisms** ☐ **Exclude exon/exon boundaries**

4. Receive result
○ via email:
● on web browser.

5.

Figure 2. This form allows you to select candidate oligos for further analysis.

neighbors and one 3′-dangling nucleotide are summed. The energy unit is K/mol at 1 M NaCl, pH 7 and 37 °C. Using the stability information one may wish to select siRNA duplexes that are less stable at the 5′ end of the anti-sense siRNA.

Because the objective of using siRNAs is to silence a specific gene in a mammalian cell, the base-pairing region for a siRNA must be selected carefully to avoid similarity to an unrelated mRNA. Therefore, our program BLASTs (12) each siRNA candidate against the human or mouse UniGene (13) database, and parses the output into an HTML table for users. Furthermore, we have developed an algorithm for mapping the siRNA to the genomic sequences and indicate to the user if the siRNA is at an exon–exon junction or if the siRNA contains any SNPs. The positions of these junctions and any SNPs are calculated in the following manner. Each week every UniGene entry is BLATted (14) against the current genome build. The position of the siRNA in the genome is calculated by mapping its position on the best-hit sequence (see below) with the genomic positions of the best hit. This position is used to calculate the exon–exon boundaries. Starting your design with either a GenBank accession number or a sequence, SNP and 3′-UTR/coding/5′-UTR locations (see below) are calculated by mapping the siRNA position on the best-hit sequence with the GenBank entry of the best hit.

There are two options for getting your final results—you can receive the URL for your results by email or you can wait for the results to appear in the browser. If you are submitting more than a dozen sequences for further analysis, it is best to receive the results by email.

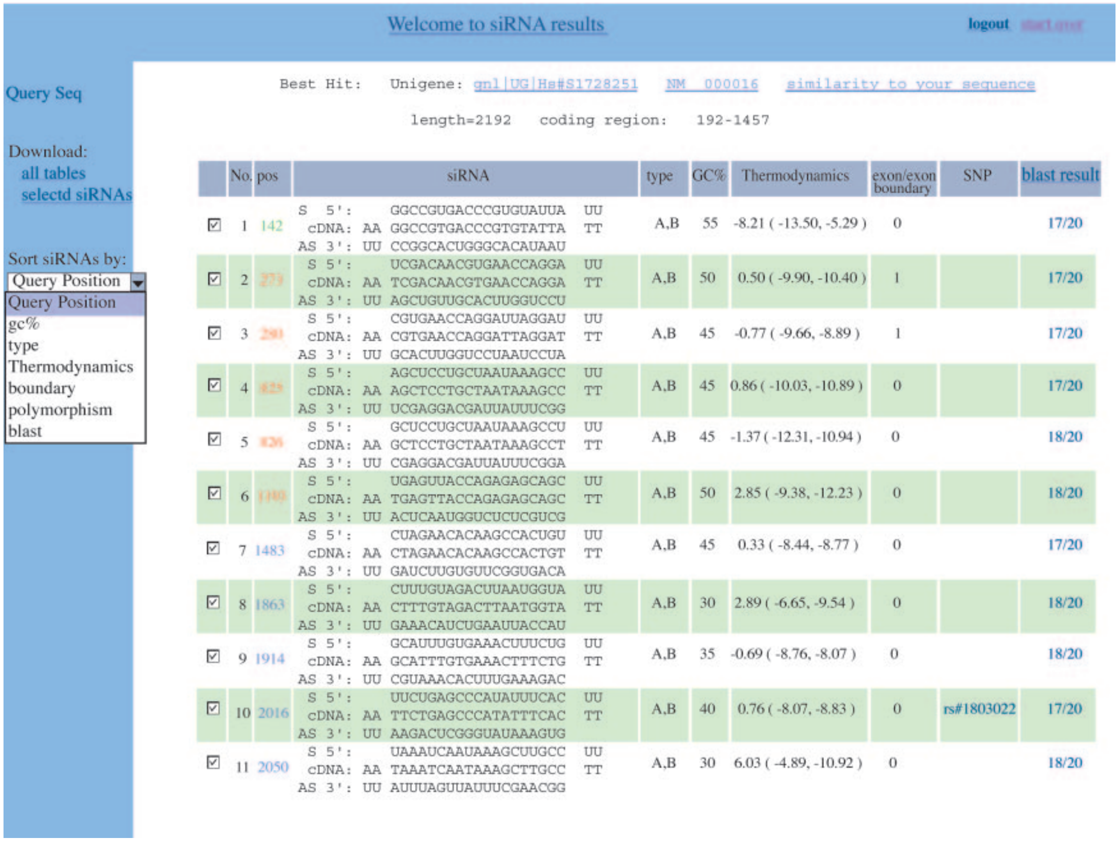
THE RESULTS PAGE

The best hit to your input sequence is found by BLASTing your query sequence against the UniGene database (Figure 3). After the BLAST search is completed, the results are displayed in tabular form with links to explore the BLAST results (Figure 4). By default, the oligos in the table are sorted by the position of the siRNA within the target sequence. With the Select box on the left menu bar, you can re-sort the siRNAs by any one of the following criteria: query position, type (siRNA pattern), GC%, thermodynamic values, exon—exon boundary, the number of SNPs and BLAST result.

The positions of the candidate oligos in the target sequence are colour coded to represent the region of the gene in which they fall. For the best UniGene hit in the candidate oligos, green indicates that the oligo is upstream of the CDS. Red indicates that the oligo is within the CDS and blue indicates that the oligo is downstream of the CDS. If the best GenBank hit is an NM entry, then the green region is the 5′-UTR and the blue indicates that the oligo is in the 3′-UTR.

DOWNLOADING DATA

There are two options for downloading the results of your siRNA design. You can choose to download all of the information in the result table or just the information about the siRNAs. Both options produce tab-delimited files that can easily be read into a spreadsheet.



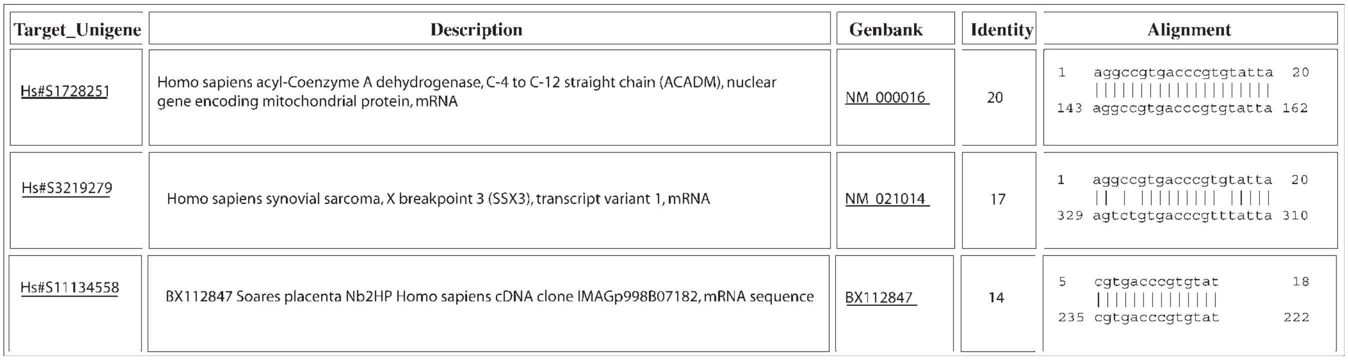
Welcome to siRNA results

Best Hit: Unigene: gnl|UG|Hs#S1728251 NM_000016 similarity to your sequence

length=2192 coding region: 192-1457

	No.	pos	siRNA	type	GC%	Thermodynamics	exon/exon boundary	SNP	blast result
☑	1	142	S 5': GGCCGUGACCCGUGUAUUA UU cDNA: AA GGCCGTGACCCGTGTATTA TT AS 3': UU CCGGCACUGGGCACAUAAU	A,B	55	-8.21 (-13.50, -5.29)	0		17/20
☑	2	273	S 5': UCGACAACGUGAACCAGGA UU cDNA: AA TCGACAACGTGAACCAGGA TT AS 3': UU AGCUGUUGCACUUGGUCCU	A,B	50	0.50 (-9.90, -10.40)	1		17/20
☑	3	290	S 5': CGUGAACCAGGAUUAGGAU UU cDNA: AA CGTGAACCAGGATTAGGAT TT AS 3': UU GCACUUGGUCCUAAUCCUA	A,B	45	-0.77 (-9.66, -8.89)	1		17/20
☑	4	825	S 5': AGCUCCUGCUAAUAAAGCC UU cDNA: AA AGCTCCTGCTAATAAAGCC TT AS 3': UU UCGAGGACGAUUAUUUCGG	A,B	45	0.86 (-10.03, -10.89)	0		17/20
☑	5	826	S 5': GCUCCUGCUAAUAAAGCCU UU cDNA: AA GCTCCTGCTAATAAAGCCT TT AS 3': UU CGAGGACGAUUAUUUCGGA	A,B	45	-1.37 (-12.31, -10.94)	0		18/20
☑	6	1193	S 5': UGAGUUACCAGAGAGCAGC UU cDNA: AA TGAGTTACCAGAGAGCAGC TT AS 3': UU ACUCAAUGGUCUCUCGUCG	A,B	50	2.85 (-9.38, -12.23)	0		18/20
☑	7	1483	S 5': CUAGAACACAAGCCACUGU UU cDNA: AA CTAGAACACAAGCCACTGT TT AS 3': UU GAUCUUGUGUUCGGUGACA	A,B	45	0.33 (-8.44, -8.77)	0		17/20
☑	8	1863	S 5': CUUUGUAGACUUAAUGGUA UU cDNA: AA CTTTGTAGACTTAATGGTA TT AS 3': UU GAAACAUCUGAAUUACCAU	A,B	30	2.89 (-6.65, -9.54)	0		18/20
☑	9	1914	S 5': GCAUUUGUGAAACUUUCUG UU cDNA: AA GCATTTGTGAAACTTTCTG TT AS 3': UU CGUAAACACUUUGAAAGAC	A,B	35	-0.69 (-8.76, -8.07)	0		18/20
☑	10	2016	S 5': UUCUGAGCCCAUAUUUCAC UU cDNA: AA TTCTGAGCCCATATTTCAC TT AS 3': UU AAGACUCGGGUAUAAAGUG	A,B	40	0.76 (-8.07, -8.83)	0	rs#1803022	17/20
☑	11	2050	S 5': UAAAUCAAUAAAGCUUGCC UU cDNA: AA TAAATCAATAAAGCTTGCC TT AS 3': UU AUUUAGUUAUUUCGAACGG	A,B	30	6.03 (-4.89, -10.92)	0		18/20

Figure 3. Result page after a BLAST search is done on candidate oligos.



Target_Unigene	Description	Genbank	Identity	Alignment
Hs#S1728251	Homo sapiens acyl-Coenzyme A dehydrogenase, C-4 to C-12 straight chain (ACADM), nuclear gene encoding mitochondrial protein, mRNA	NM 000016	20	1 aggccgtgacccgtgtatta 20 |||||||||||||||||||| 143 aggccgtgacccgtgtatta 162
Hs#S3219279	Homo sapiens synovial sarcoma, X breakpoint 3 (SSX3), transcript variant 1, mRNA	NM 021014	17	1 aggccgtgacccgtgtatta 20 || | ||||||||| ||||| 329 agtctgtgacccgtttatta 310
Hs#S11134558	BX112847 Soares placenta Nb2HP Homo sapiens cDNA clone IMAGp998B07182, mRNA sequence	BX112847	14	5 cgtgacccgtgtat 18 |||||||||||||| 235 cgtgacccgtgtat 222

Figure 4. Summary BLAST results for each candidate siRNA.

FUTURE DEVELOPMENTS

We are continuing to develop our website to include additional capabilities. One such feature is to allow the user to design siRNA of various lengths rather than only 21mers. In addition we will be adding a batch capability so that multiple sequences can be input at once. Our site will soon provide assistance in predicting short hairpin siRNAs. Also, as new rules are discovered by our group and others, we will incorporate them into our siRNA design server.

SUMMARY

In summary, we wish to emphasize that our siRNA design server was built in collaboration with laboratory scientists. Our main goal is to provide the most accurate siRNA results and flexibility for the end user so that resulting information can be effectively evaluated.

ACKNOWLEDGEMENTS

We wish to acknowledge Brent Stockwell, Whitehead Institute for suggesting that we build this server, Carl Novina, MIT for comments and suggestions, and Tom DiCesare, Whitehead Institute for graphics help.

REFERENCES

1. Dykxhoorn,D.M., Novina,C.D. and Sharp,P.A. (2003) Killing the messenger: short RNAs that silence gene expression. *Nat. Rev. Mol. Cell. Biol.*, **4**, 457–467.
2. Elbashir,S.M., Harborth,J., Lendeckel,W., Yalcin,A., Waber,K. and Tuschl,T. (2001) Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature*, **411**, 494–498.
3. Elbashir,S.M., Harborth,J., Weber,K. and Tuschl,T. (2002) Analysis of gene function in somatic mammalian cells using small interfering RNAs. *Methods*, **26**, 199–213.
4. Harborth,J., Elbashir,S.M., Vandenburgh,K., Manninga,H., Scaringe,S.A., Weber,K. and Tuschl,T. (2003) Sequence, chemical, and structural variation of small interfering RNAs and short hairpin RNAs and the effect on mammalian gene silencing. *Antisense Nucleic Acid Drug Dev*., **13**, 83–105.
5. Tuschl,T. (2002) Expanding small RNA interference. *Nat. Biotechnol*., **20**, 446–448.
6. Shi,Y. (2003) Mammalian RNAi for the masses. *Trends Genet*., **19**, 9–12.
7. Schwarz,D.S., Hutvagner,G., Du,T., Xu,Z., Aronin,N. and Zamore,P.D. (2003) Asymmetry in the assembly of the RNAi enzyme complex. *Cell*, **115**, 199–208.
8. Khvorova,A., Reynolds,A. and Jayasena,S.D. (2003) Functional siRNAs and miRNAs exhibit strand bias. *Cell*, **115**, 209–216.
9. Aza-Blanc,P., Cooper,C.L., Wagner,K., Batalor,S., Deveraux,Q.L. and Cooke,M.P. (2003) Identification of modulators of TRAIL-induced apoptosis via RNAi-based phenotypic screening. *Mol. Cell*, **12**, 627–637.
10. Reynolds,A., Leake,D., Boese,Q., Scaringe,S., Marshall,W.S. and Khvorova,A. (2004) Rational siRNA design for RNA interference. *Nat. Biotechnol*., **22**, 326–330.
11. Mathews,D.H., Sabina,J., Zuker,M. and Turner,D.H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol*., **288**, 911–940.
12. Altschul,S., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipmen,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.
13. Wheeler,D.L., Church,D.M., Edgar,R., Ferderhen,S., Helmberg,W., Madden,T.L., Pontius,J.U., Schuler,G.D, Schriml,L.M., Sequeira,E. *et al*. (2004) Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res*., **32**, D35–D40.
14. Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res*., **12**, 656–664.